Jetstream and Brain-Life: Creating a reproducible research platform on the Jetstream cloud

STEVEN O'RILEY, Indiana University
AMAN ARYA, University of Washington

Jetstream is the first production cloud funded by the National Science Foundation (NSF) for conducting general science and engineering research. Brain-Life is an in-development grassroots research platform with the proposed goal to provide an easy-to-use platform for Neuroscientists and other STEM (Science, Technology, Engineering, Mathematics) researchers to publish their data and algorithms, as well as manipulate archived data with on-demand compute resources.

One of the core compute resources being used by the Brain-Life platform is the Jetstream production cloud. In this paper, we discuss the ease by which the Brain-Life platform is able to integrate itself with the Jetstream cloud and depict several applications on Brain-Life that have been successfully implemented using the Jetstream cloud service. We also identify issues in the integration of Brain-Life and Jetstream, and the problems which show up when trying to run algorithms on the Jetstream cloud via the Brain-Life platform.

Additional Key Words and Phrases: Jetstream, Brain-Life, Reproducibility, Data Storage, Science Gateway, Clusters, HPC Systems

ACM Reference format:
Steven O'Riley and Aman Arya. 2017. Jetstream and Brain-Life: Creating a reproducible research platform on the Jetstream cloud. 1, 1, Article 1 (August 2017), 3 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

One of the great strengths of modern scientific research is its being a transdisciplinary field that requires engineers, scientists, mathematicians, and statisticians to work towards a common goal. 'Open' data sharing and scientific cooperation dissolve disciplinary boundaries and allow STEM researchers to use methods from traditionally different fields in their own research. However, in order to ensure the success of this process, the problem of establishing standardized mechanisms which assert the reproducibility of research must first be overcome.

Brain-Life provides users with an easy-to-use browser-based research platform built around publishing data, running common algorithms used in Neuroscience and other STEM fields, and providing the high end computing resources which are used to manage and manipulate archived data.

The Brain-Life platform also includes a highly configurable network configuration which natively supports the Jetstream cloud, [Overview 2017] a service which is built for research in the "long-tail-of-science." [Fischer et al. 2017] Jetstream has already been utilized for a number of other science gateways, including Galaxy and SEA-Grid [Knepper et al. 2017], and Brain-Life is a welcome addition to this list. As a Neuroscience gateway, the Brain-Life platform serves as a centralized location where brain data and its derivatives can be stored, manipulated, and shared in a standardized way.

In this paper, we discuss how Brain-Life is configured and integrated with the Jetstream cloud service, as well as how a specific set of use-case algorithms developed on the platform illustrate the fluidity of this integration. We discuss the pros and cons of using Jetstream versus other cloud services, as well as planned revisions for the future of the Jetstream-BrainLife architecture.

2 JETSTREAM-BRAINLIFE ARCHITECTURE

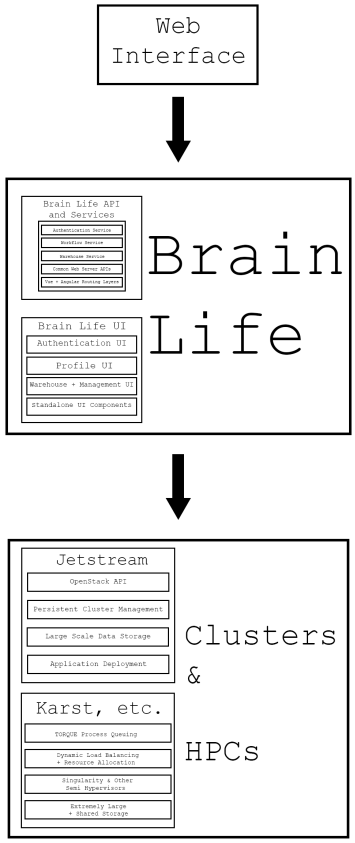

Fig. 1. Brain-Life architecture

Authors' Addresses: Steven O'Riley, Department of Computer Science, Indiana University, 1320 E 10th St, Bloomington, IN 47405, US; Aman Arya, Applied & Computational Mathematical Sciences, University of Washington, West Stevens Way Northeast, Seattle, WA 98105, US.

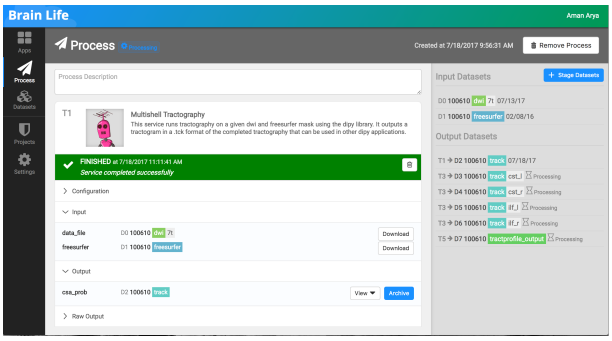

Fig. 2. Sample Brain-Life Process

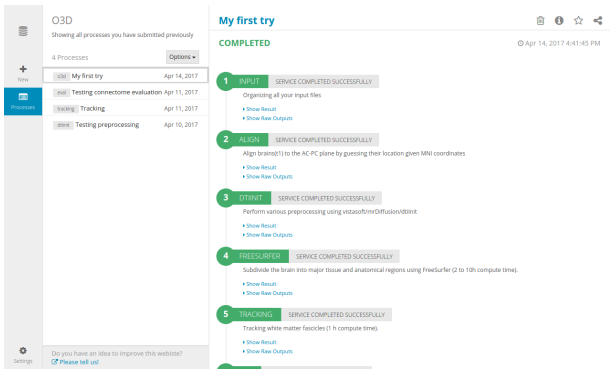

Fig. 3. Sample Brain-Life Pipeline

2.1 Topology

As a web portal, Brain-Life is able to connect the end user to High Performance Computing (HPC) systems and Virtual Machines (VMs) through the use of a friendly-to-use User Interface (UI). Its UI has been designed in such a way that the overall aesthetics and components that make up the front end of the website can be modified separately from the backend Application Programming Interface (API) that converses with larger Cyberinfrastructure. Brain-Life utilizes these systems for frequent storage of big data, namely from the Human Connectome Project (HCP). Data can be manipulated through the use of modularized algorithms in the form of applications, each of which are open sourced on GitHub. [Overview 2017]

2.2 Brain-Life Applications

Applications on Brain-Life serve as self-contained algorithms which are easily accessible via a user interface provided by the Brain-Life platform. The user can pick and choose datasets which are available from a large range of resources, including HCP, or upload their own data and run a set of processes on them (Fig 3). Users can additionally combine multiple processes into a single, reusable pipeline (Fig 2).

Current applications that run on Brain-Life via Jetstream are Multishell Tractography, White Matter Segmentation, Tract Profiling, and CSA (Constant Solid Angle) Peaks. More applications are constantly being developed to be used by Jetstream and other cloud services. [Overview 2017]

It is very simple for users to implement their own applications to be used and referenced later in derivative research papers. The creator of each application is able to use whatever programming language they are most comfortable with to interface with Brain-Life's API.

2.3 High End Cyberinfrastructure

Through the utilization of a browser-based UI and a set of apps, users are able to manipulate and store datasets to the cloud, the final component of Brain-Life's designed architecture. A cloud can represent either a homogeneous cluster such as Jetstream or an HPC system like Indiana University's Karst. The cloud also serves as the location where apps are cloned, run, and evaluated. Each app is flexibly designed so that it can run either on an active VM or an HPC system, and future plans for Brain-Life include dockerizing each application in order to elasticize this flexibility.

Moreover, datasets can be either directly uploaded or archived to the cloud after each process pipeline has completed. And, each dataset is stored in accordance with the Brain Imaging Data Structure (BIDS) to further standardize the way data can be utilized and reproduced later. When users have completed their pipeline of tasks, they can download the resulting data from the cloud and reference it by Brain-Life's auto-generated data object identifier (DOI) in their research papers. [Overview 2017]

3 ADVANTAGES OF JETSTREAM

Some advantages of using Jetstream over other HPC systems such as Karst include the following: [Stewart et al. 2016]

- Applications run without waiting in queues (as opposed to other HPC systems, which utilize TORQUE)
- Jetstream is more powerful and is generally faster at running most applications on Brain-Life
- Jetstream provides root access to all VMs
- The architecture of Jetstream allows for users to continue their research, even while maintenance may occur (as opposed to HPC systems, where a period of maintenance requires users to stop researching for days)

Jetstream importantly provides users with the ability to run their applications without being hindered by wall times or queues, as opposed to HPC systems. Research is able to be conducted continuously and without unnecessary downtime. Providing root access to all VMs allows researchers to easily configure their environments by themselves without having to rely on a shared, static filesystem, which reduces time and frustration for both users and administrators. Furthermore, constant maintenance can be a hindrance to research efforts, and Jetstream's ability to allow researchers to continue their studies, even on maintenance days, removes this obstacle altogether.

4 IMPROVEMENTS YET TO BE MADE

While the use of Jetstream has managed to solve previous problems such as lengthy computational runtime and continuous reliability, there are a large number of improvements that can be made to

Brain-Life. It has been previously mentioned that the platform will soon employ an application system based on dockerization so that issues of dependencies and reusability (especially on HPC systems) can be solved almost entirely. Another planned improvement is modularizing the API components that make up Brain-Life across different machines (i.e. dataset storage, process task handling, metadata management) which can each be individually improved and scaled.

There remain other micro-modifications and improvements which can be made to different portions of Brain-Life's architecture, but overall the goal of each revision is to build towards serving the original purpose of the platform's existence: to provide a centralized location where data from various STEM fields, namely Neuroscience, can be uploaded, shared, modified, reproduced, and downloaded in a way that is standardized and which will drastically increase the speed at which new research can be done, and frequent the growth of science overall. [Overview 2017]

5 CONCLUSION

Brain-Life has accelerated in its progress to become a platform whereupon data and algorithms can be freely shared, utilized, and reproduced. And a large portion of its recent success can be directly attributed to the integration of the Jetstream cloud.

Jetstream offers a unique architecture and powerful advantages over other queue-based HPC systems like Indiana University's Karst, which may not always be available. The combination of its architecture and usability, designed for the long-tail of scientific research, makes the Jetstream cloud an ideal choice for integration with Brain-Life. Overall, Jetstream has been a perfect match for the type of on-the-fly large-scale STEM-related data analysis that Brain-Life has sought to offer to transdisciplinary researchers from when it was first conceived.

REFERENCES

Jeremy Fischer, George Turner, David Y Hancock, Winona Snapp-childs, John Michael Lowe, and Craig A Stewart. 2017. Jetstream : A cloud system enabling learning in higher education communities. (2017).

Richard Knepper, Eric Coulter, Marlon Pierce, Suresh Marru, and Sudhakar Pamidighantam. 2017. Using the Jetstream Research Cloud to provide Science Gateway resources. (2017). https://doi.org/10.1109/CCGRID.2017.121

Training Overview. 2017. Franco's proposal: we will develop a platform to capture brain data, publish algorithms as reproducible applications, and perform data-intensive computing on high- performance compute clusters, NSF-funded public and commercial clouds. âĂŃ. (2017), 1–23.

Craig A Stewart, David Y. Hancock, Matthew Vaughn, Jeremy Fischer, Lee Liming, Nirav Merchant, Therese Miller, John Michael Lowe, Daniel Stanzione, Jaymes Taylor, and Edwin Skidmore. 2016. Jetstream - Performance, Early Experiences, and Early Results. In *Proceedings of the XSEDE16 Conference*. St. Louis, MO. https://doi.org/10.1145/2949550.2949639